VIDEO ANALYSIS AND SYNTHESIS BASED ON A RETINAL-INSPIRED FRAME

Effrosyni Doutsi[1,2], *Lionel Fillatre*[1], *Marc Antonini*[1] *and Julien Gaulmin*[2]

[1]Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France
[2]4G-TECHNOLOGY, 460 avenue de la Quiera 06370 Mouans Sartoux - France.

ABSTRACT

This paper introduces a novel retinal-inspired filter which is applied on video streams. We mathematically prove that under specific assumptions the spatiotemporal convolution turns into a spatial convolution with a short lifespan temporal kernel. As a consequence, the filter is applied on each image of the video stream separately. We analyze how each image is decomposed into a group of subbands, each one of which approximates the image providing different kind of information. Afterwords, we propose an algorithm to reconstruct each image by exploiting the group of subbands. Finally, we defend our mathematical proofs by providing numerical simulations which show the relevance of our study.

Index Terms— Retinal-inspired processing, non-separable spatiotemporal filter, frame theory, dual frame.

1. INTRODUCTION

The research related to image and video compression algorithms remains one of the most challenging scientific fields. This is due to the fact that images and videos are widely utilised not only for personal use but also for security reasons. As a result, a big amount of data need to be transmitted and/or saved in real time satisfying multiple constraints. These constraints are mostly related to the network bandwidth, the memory of the system, the distortion of the data or the energy of the system. The combination of all these constraints would give an optimal solution but, in practice, one should seek for a relevant trade-off between them.

Closed-Circuit TeleVision systems (CCTV) are one of the video processing applications which have been involved in the exponential increase of data. The most important challenge in this kind of systems is to minimize the energy consumption of the system which is totally related to the compression rate and the bandwidth of the real time transmision. At the same time, whatever the bandwidth is, it is always necessary to transmit only the most informative and meaningful data such that the reconstruction quality will be the best possible one. As a result, it would be advantageous for CCTV system if we could propose an algorithms which saves power.

Trying to deal with the above problem we got inspired by the visual system in order to propose an alternative decomposition model for video. The retinal function has been explicitly modeled by neuroscientists and the experimental results have shown that this should be an efficient "compression" model especialy with respect to the energy minimization [1, 2]. This is due to the fact that the retina is a layered structure of different kinds of cells, the amount of which decreases while they turn to connect to the optic nerve [3].

In this paper, our goal is to study the retinal-inspired transformation from the signal processing point of view in order to save power and set it as basis for our future bio-inspired dynamic codec. The first attempt in modeling this kind of filter was proposed as a bio-inspired codec of natural images by [4]. The authors tried to approximate the spatiotemporal variations of the retinal filtering using a Difference of Gaussians (DoG) pyramid based on [5] and [6], considering at the same time that each layer appears at different moment according to an exponential temporal function. We improve this filter by taking into account explicitly the time in the design of our novel *non-Separable sPAtioteMporal (non-SPAM)* filter. The advantage of the non-SPAM filter is the fact that when a stimulus appears its non-SPAM transformation is based not only on the spatial neighborhood for the given time but also on previous times. As a result, this is a spatiotemporal transformation which enables to enrich the details of the signal.

In section 2, we introduce the non-SPAM filter and explain its bio-inspired nature. Then, in section 3, we introduce how the filter is able to decompose the input video. We prove that the non-SPAM filter is invertible in section 4. In section 5, we propose the non-SPAM synthesis based on the frame theory. The numerical results are given in section 6. Section 7 concludes the paper.

2. OBJECTIVE AND RETINAL MODEL

The general aim of this study is to introduce a novel codec for videos, captured for surveillance and/or security reasons, which need to be transmitted through the network to a client who is going to display and analyze the scenes (Fig. 1). The variations of the network bandwidth which depend on the location of the captured area are combined to the complexity of the scene. As a result, it is necessary to built a special archi-

Thanks to 4G-TECHNOLOGY and ANRT for funding.

tecture which stands as a trade-off between them. For that reason, we have been inspired by the visual system which codes the luminance of light that reaches the eyes and transubstantiate them into spike trains (electrical impulses) which include all the necessary information of the input signal. It seems that this kind of code is efficient enough in order to be used in the reconstruction of the signal which is necessary in image processing.

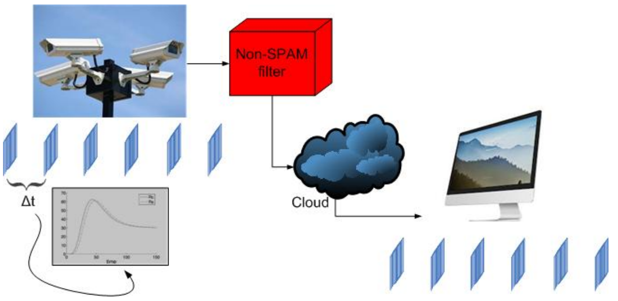


Fig. 1: Non-SPAM compression schema. A video stream is captured by a CCTV system. Each image is filtered by the non-SPAM in order to be transmitted to the user where it is decoded and displayed. The time Δt between two images equals the life-time of the temporal filters $R_c(u)$ and $R_s(u)$ which stand for the temporal behavior of the non-SPAM and tune the spatial changes of the filter.

Our primary goal is to mimic the anatomy of retina and the functions of each group of the retinal cells in terms of signal processing. Based on [2, 7] we assume that the group of cells which form the outer plexiform layer (photoreceptors, horizontal and biopolar cells) receives the light and spatially decomposes the signal with respect to their sensitivity into blurred versions. Each of these blurred versions is temporally enriched in details while the signal is transmitted on the way to ganglion cells. These are the features that our filter, the *non-Separable sPAtioteMporal (non-SPAM)* filter, tries to mimic having a spatial behavior which varies with respect to time.

2.1. Retinal Model

We consider that a video consists of N different images. Each one of which is generated in specific time g_i and lasts until time g_{i+1} when the next image appears. Throughout the paper, we consider that a video is composed of images instead of frames in order to avoid any confusion with the frame theory vocabulary. Let us define a video in continuous time:

$$V(x,t) = \sum_{i=1}^{N} f_i(x)\mathbb{1}_{[g_i,g_{i+1}]}(t), \tag{1}$$

where $x \in \mathbb{R}^2$, $t \in [0,T]$, $T \in \mathbb{R}^+$ is the length of the video, $f_i(x)$ stands for the i-th image of the video, N is the total number of images which form the video stream and $\mathbb{1}_{[g_i,g_{i+1}]}(t)$ is the indicator function which equals to 1 if $g_i \le t \le g_{i+1}$, and 0 otherwise. The ideal spatiotemporal convolution of the non-SPAM and the video results in the function $A(x,t)$ which is called the *activation degree* is defined by:

$$A(x,t) = K(x,t) \overset{x,t}{*} V(x,t) \tag{2}$$

where $\overset{x,t}{*}$ is the convolution with respect to space and time. The non-SPAM filter mimics the function of the outer plexiform layer of the retina. The space stands for the spatial transformation of the receptors and the time for temporal improvement of the initial transform by the center-suround structure of horizontal and bipolar cells [7]. With this filter we obtain a retinal-inspired image decomposition instead of the convential ones i.e DCT [8], DWT [9] or filter banks [5]. Based on [2], we define the non-SPAM filter in continuous time and space, as:

$$K(x,t) \quad = C(x,t) - S(x,t), \tag{3}$$

where $C(x,t)$ and $S(x,t)$ are the center and the surround spatiotemporal filters given by (4) and (5) respectively:

$$C(x,t) \quad = w_c G_{\sigma_C}(x) W(t), \tag{4}$$

$$S(x,t) \quad = w_s G_{\sigma_S}(x) \left(W \overset{t}{*} E_{\tau_S} \right)(t), \tag{5}$$

where w_c and w_s are constant parameters, G_{σ_C} and G_{σ_S} are spatial Gaussian filters standing for the center and surround areas respectively, and E_{τ_S} is an exponential temporal filter. The center temporal filter $W(t)$ is given by:

$$W(t) = E_{\tau_G,n} \overset{t}{*} \left(\delta_0 - w_c E_{\tau_C}\right)(t), \tag{6}$$

where the gamma temporal filter $E_{\tau_G,n}(t)$ is defined by:

$$E_{\tau,n}(t) = \frac{t^n \exp\left(-t/\tau\right)}{\tau^{n+1}}, \tag{7}$$

with $n \in \mathbb{N}$, τ is a constant parameter ($E_{\tau,n}(t) = 0$ for $t < 0$), E_{τ_C} is an exponential temporal filter, δ_0 is the dirac function and $\overset{t}{*}$ stands for the temporal convolution. In case that $n = 0$, the gamma filter turns to an exponential filter. The convolution of the temporal filter $W(t)$ with the exponential filter $E_{\tau_S}(t)$ is related to the delay in the appearance of the surround temporal filter with respect to the center one.

3. RETINAL MODEL ANALYSIS

The calculation of the activation degree $A(x,t)$ in (2) applied to the video $V(x,t)$ in (1) turns into a spatial convolution with a time-varying kernel as proved in the following proposition. To simplify the calculation, it is assumed that $g_{i+1} - g_i = \Delta t$ is constant for all $i = 1, \ldots, N$.

Proposition 1. *For all $t < g_1$, the activation degree $A(x,t)$ in (2) applied on $V(x,t)$ in (1) is $A(x,t) = 0$. For $t \geq g_1$, $A(x,t)$ can be rewritten as:*

$$A(x,t) = \sum_{i=1}^{N} \phi(x, t-g_i) \overset{x}{*} f_i(x), \tag{8}$$

where $\phi(x,u)$ is a spatial DoG filter weighted by two temporal filters $R_c(u)$ and $R_s(u)$ satisfying:

$$\phi(x,u) = w_c G_{\sigma_C}(x) R_c(u) - w_s G_{\sigma_S}(x) R_s(u), \tag{9}$$

$$R_c(u) = \int\limits_{\max\{0, u-\Delta t\}}^{u} W(\ell) d\ell, \tag{10}$$

$$R_s(u) = \int\limits_{\max\{0, u-\Delta t\}}^{u} W(\ell) \overset{\ell}{*} E_{\tau_s}(\ell) d\ell, \tag{11}$$

and $R_c(u) = R_s(u) = 0$ for $u < 0$.

The proof of Proposition 1 is omitted due to the lack of place. According to Proposition 1, the activation degree $A(x,t)$ depends on all the images $f_i(x)$ occurring before time t but the following corollary shows that, under some mild assumptions, the activation degree can be processed image per image. Before presenting this corollary, let us introduce a useful lemma.

Lemma 1. *The function $\phi(x,u)$ is a continuous and infinitely differential function for all $u \geq 0$ and all $x \in \mathbb{R}^2$ such that:*

$$\lim_{u \to +\infty} \phi(x,u) = \phi(x), \forall x \in \mathbb{R}^2, \tag{12}$$

where $\phi(x)$ is a DoG filter independent of u.

Lemma 1 shows that $\phi(x,u)$ converges toward a constant spatial DoG filter as u tends to infinity. This convergence is illustrated in Fig. 2. Let $\varepsilon > 0$ be a small positive constant. According to Lemma 1, there exists $t_c = t_c(\varepsilon) > 0$ such that

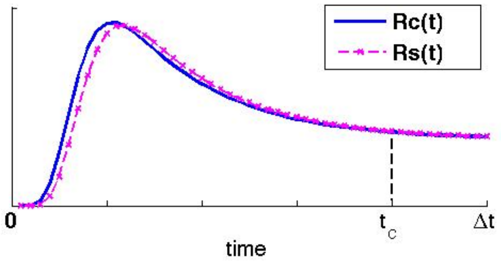


Fig. 2: Temporal filters $R_c(u)$ and $R_s(u)$.

$$|\phi(x,u) - \phi(x)| < \varepsilon, \forall u \geq t_c, \forall x \in \mathbb{R}^2. \tag{13}$$

The following corollary comes from Proposition 1 together with Lemma 1.

Corollary 1. *Let $\varepsilon > 0$ and assume that the parameters of $\phi(x,u)$ are chosen such that $t_c(\varepsilon) < \Delta t$. Let t such that $g_1 \leq t \leq g_{N+1}$ and i be the unique integer such that $g_i \leq t < g_{i+1}$. Then, the activation degree $A(x,t)$ in (8) can be approximated by*

$$\widehat{A}(x,t) = f_i(x) \overset{x}{*} \phi(x, t-g_i) + \sum_{j:\ g_j + t_c < t} f_j(x) \overset{x}{*} \phi(x) \tag{14}$$

where $|\widehat{A}(x,t) - A(x,t)| < \eta$ with η a small positive constant directly proportional to ε. It follows that:

$$\widehat{A}(x,t) = A_i(x,t) + B_i(x) \tag{15}$$

where $A_i(x,t)$ is the filtered version of $f_i(x)$:

$$A_i(x,t) = \phi(x, t-g_i) \overset{x}{*} f_i(x), \tag{16}$$

and $B_i(x)$ is defined recursively by $B_1(x) = 0$ and

$$B_{i+1}(x) = B_i(x) + A_i(x) \tag{17}$$

with

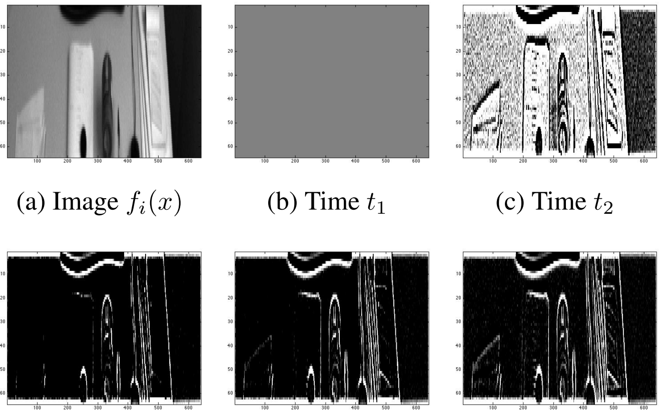
$$A_i(x) = \phi(x) \overset{x}{*} f_i(x).$$

The interest of Corollary 1 is to show that, at time t, the activation degree $A(x,t)$ can be approximated by $\widehat{A}(x,t)$ which only depends on t via $A_i(x,t)$. The remaining term $B_i(x)$ in (15) corresponds to the $A_j(x,t)$'s for $j < i$ occurring before time g_i in (8). Since Lemma 1 yields $A_j(x,t_c) \approx A_j(x)$, the remaining term $B_i(x)$ is (almost) time independent and it does not convey any information on $f_i(x)$. The factor $A_j(x)$ is transmitted at the end of time interval $[g_j, g_{j+1}]$, hence, in practice, it is not useful to compute the full convolution $\widehat{A}(x,t)$ in (15). It is sufficient to apply the non-SPAM filter on image $f_i(x)$ during the time interval $[g_i, g_i + t_c]$ and to transmit $A_i(x,t)$.

The above corollary is crucial for the reason that it enables the simplification and representation of the non-SPAM filter like a block of time-varying DoG kernels. The DoG kernels have been extensively studied in the past [6, 10, 11] and they can be processed efficiently. Fig. 3 shows the non-SPAM decomposition of one image, say $f_i(x)$, which is extracted from a video stream. We have selected 5 different time samples $t \in \{t_1, t_2, t_3, t_4, t_5\}$ of $A_i(x,t)$ where $g_i \leq t_j \leq g_{i+1}$.

4. NON-SPAM FRAME

The goal of this section is to recall that the non-SPAM filter, when it is applied on each image separately, is invertible and permits us to reconstruct the video image per image. For this reason, we establish that the non-SPAM filter has a frame structure according to the frame theory [11, 12]. Let us consider the image $f_i(x)$ over the time interval $[g_i, g_{i+1}]$. As underlined in the discussion following Corollary 1, the decoder



(a) Image $f_i(x)$ (b) Time t_1 (c) Time t_2

(d) Time t_3 (e) Time t_4 (f) Time t_5

Fig. 3: Non-SPAM filter applied to $f_i(x)$ at 5 time samples.

receives a stream of images $A_i(x,t)$ described by (16) and its goal is to reconstruct $f_i(x)$ from $A_i(x,t)$ for $t \in [g_i, g_{i+1}]$.

For numerical purpose, we need to discretize the non-SPAM filter in space and in time. Let $x_1, \ldots, x_n \in \mathbb{R}^2$ be some sets of spatial sampling points and $t_1, \ldots, t_m \in [g_i, g_{i+1}]$ be temporal sampling points. Let us denote $u_1 = t_1 - g_i, \ldots, u_m = t_m - g_i$ be the elapsed times between the t_j's and g_i. Without any loss of generality, it is assumed that the u_i's are the same whatever the considered time interval $[g_i, g_{i+1}]$. As a consequence, the continuous spatial convolution $A_i(x,t)$ is approximated by the discrete convolution:

$$\begin{aligned} A_i(x_k, t_j) &= \phi(x_k, t_j - g_i) \circledast f_i(x_k) \\ &= \sum_{p=1}^{n} \phi(x_k - x_p, u_j) f_i(x_p) = A_i(x_k, u_j), \end{aligned}$$

for all $1 \leq k \leq n$ and $1 \leq j \leq m$. Let $\varphi_{k,j}$ be the row vector of $\mathbb{R}^n$ defined by

$$\varphi_{k,j} = \Big(\phi(x_k - x_1, u_j), \ldots, \phi(x_k - x_n, u_j)\Big). \qquad (18)$$

Let us denote f_i the sampled version of the image $f_i(x)$:

$$f_i = \left(f_i(x_1), \ldots, f_i(x_n)\right), \qquad (19)$$

and $\|f_i\|$ be the Euclidean norm of f_i. In our previouw work, we have proven that the family of vectors Φ is a frame [13,14].

5. RETINAL MODEL SYNTHESIS

The optimal reconstruction of the input video, image per image, is possible when we provide to the decoder all the coefficients of the non-SPAM image. At time t_m ending the interval $[g_i, g_{i+1}]$, the optimal estimate $\hat{f}_i$ of f_i is given by:

$$\hat{f}_i = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \underline{A_i}, \qquad (20)$$

where M^{-1} denotes the inverse of a matrix M and $M^\top$ denotes its transpose. With a short abuse of notation, Φ is a family of vector of size the $nm \times n$ given by $\Phi = \varphi_{k,j} : 1 \leq k \leq n, 1 \leq j \leq m$, $\underline{A_i}$ is another vector which is given by $\underline{A_i} = A_i(x_k, u_j) : 1 \leq k \leq n, 1 \leq j \leq m$. The dual frame, which is necessary to have a perfect decoding at time t_m [11, 12], is $(\Phi^\top \Phi)^{-1} \Phi^\top$. Instead of computing the above matrix operator which is time consuming and resource demanding, we can easily shown that (20) is a solution of the following least squares problem:

$$\hat{f}_i = \arg \min_{f_i \in \mathbb{R}^n} \left(\sum_{j=1}^{m} \|\phi_j \circledast f_i - A_{i,j}\|^2 \right), \qquad (21)$$

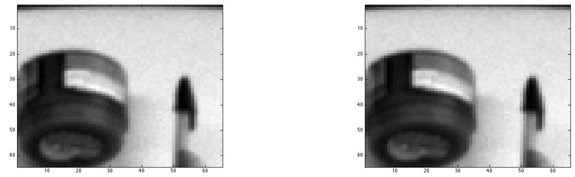
where the vectors ϕ_j and $A_{i,j}$ are defined by:

$$\begin{aligned} \phi_j &= \big(\phi(x_1, u_j), \ldots, \phi(x_n, u_j)\big), & (22) \\ A_{i,j} &= \big(A_i(x_1, u_j), \ldots, A_i(x_n, u_j)\big). & (23) \end{aligned}$$

Hence, the estimate $\hat{f}_i$ is computed by using a gradient descent algorithm.

6. NUMERICAL RESULTS

We have captured a video with a rate of 20 images per second and we have applied the non-SPAM on each image of size 64×64 using the software tool MATLAB.



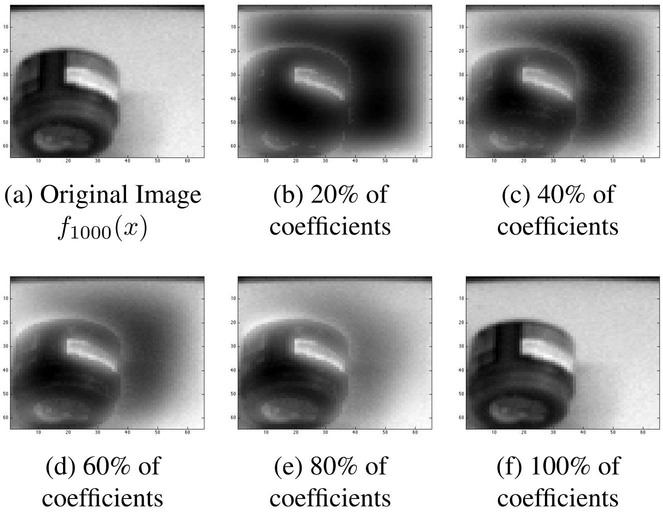
(a) Original Image (b) Reconstructed Image

Fig. 4: Reconctruction of the 400^{th} image of the video stream.

All the parameters which are related to the lifespan of the non-SPAM filter are tuned according the video rate (see Corollary 1) $\Delta t = 50$msec, $\tau_C = 10. * 10^{-3}$sec, $\tau_S = 9. * 10^{-4}$sec, $\tau_G = 1. * 10^{-3}$sec, $n = 5$, $w_c = 0.75$, and $w_s = 1$. The rest of the parameters which are related to the spatial domain, $\sigma_c, \sigma_s, w_c, w_s$, are biologically plausible and they have been obtained by modeling the center-surround structure of the retinal cells' receptive fields [1, 11].

The reconstruction results generated by the total number of coefficients are almost perfect and they are illustrated in Fig. 4. Hence, there is some redundancy within the transmitted coefficients. For this reason, we propose to apply the Rank-Order-Coding (ROC) model proposed in [1] . The ROC model is traditionally used to convert an analog signal into a rank order of electrical impulses (spikes). The spike which is first emitted has been caused by a rapid excitation because of a strong signal.

In this study, we apply the ROC model only with respect to the informative coefficients which are generated by the non-SPAM transformation but we do not aim to produce any spikes. We sort in a descending order the coefficients of each decomposition layer and we select a percentage of coefficients in A_i keeping the ones with the highest energy omitting the rests. In Fig. 5, we have selected one image of the video stream and we have decided to use 5 different percentages of the highest values of the activation degree of 5 DoG kernels.



(a) Original Image $f_{1000}(x)$

(b) 20% of coefficients

(c) 40% of coefficients

(d) 60% of coefficients

(e) 80% of coefficients

(f) 100% of coefficients

Fig. 5: Reconstruction based on the ROC model.

7. CONCLUSION

This paper proposes to study the analysis and the synthesis of a retinal-inspired filter which is applied on video streams. We have shown that this filter can be applied image per image without any significant loss of information. Our future goal is to adapt this filter into a bio-inspired codec which is going to produce an event-based code.

8. REFERENCES

[1] R. Van Rullen and S. J. Thorpe, “Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex,” *Neural Computation*, vol. 13, pp. 1255–1283, 2001.

[2] A. Wohrer and P. Kornprobst, “Virtual retina: A biological retina model and simulator, with constrast gain control.,” *Journal of Computational Neuroscience*, vol. 26, no. 2, pp. 219–249, 2009.

[3] R. Masland, “The fundamental plan of the retina.,” *Natural Neuroscience*, vol. 4, no. 9, pp. 877–886, 2001.

[4] K. Masmoudi, M. Antonini, and P. Kornprobst, “Streaming an image through the eye: The retina seen as a dithered scalable image coder,” *Signal processing Image Communication*, vol. 28, no. 8, pp. 856–869, 2013.

[5] P. Burt and E. Andelson, “The Laplacian pyramid as a compact image code,” *IEEE Trans. Commun*, vol. 31, no. 4, pp. 532–540, 1983.

[6] David J. Field, “What is the goal of sensory coding?,” *Neural Computation*, vol. 6, pp. 559–601, 1994.

[7] Helga Kolb, “How the retina works : Much of the construction of an image takes place in the retina itself through the use of specialized neural circuits,” *American Scientist the magazine of Sigma Xi, The Scientific Research Society*, vol. 91, pp. 28–35, 2004.

[8] V. Britanak, *The Transform and Data Compression Handbook -Discrete Cosine and Sine Transforms*, CRC Press LLC, 2001.

[9] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2^{nd} edition, 1999.

[10] G. DeAngelis D. Cai and R. Freeman, “Spatiotemporal receptive field organization in the Lateral Geniculate Nucleus of cats and kittens,” *The American Physiological Society*, vol. 22, no. 3077, pp. 1045–1061, 1997.

[11] K. Masmoudi, M. Antonini, and P. Kornprobst, “Frames for exact inversion of the rank order coder,” *IEEE Transactin on Neural Networks*, vol. 23, no. 2, pp. 353–359, February 2012.

[12] J. Kovacevic and A. Chebina, “An introduction to frames,” *Signal Processing*, vol. 2, no. 1, pp. 1–94, 2008.

[13] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, “Retina-inspired filtering for dynamic image coding,” *IEEE International Conference in Image Processing (ICIP)*, 2015.

[14] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, “Event-based coding of images using a bio-inspired frame,” *Internation Conference on Event-Based Control, Communication and Signal Processing (EBCCSP)*, 2015.